A Dialogue-Based Learning Analytics Framework for Collaborative Game-Based Learning

Yeo Jin Kim[1], Daeun Hong[2], Tianshu Wang[1], Wookhee Min[1]
Snigdha Chaturvedi[3], Cindy E. Hmelo-Silver[2], James Lester[1],
[1]North Carolina State University, USA
[2]Indiana University, USA
[3]UNC Chapel Hill, USA
ykim32@ncsu.edu, dh32@iu.edu, twang43@ncsu.edu, wmin@ncsu.edu,
snigdha@cs.unc.edu, chmelosi@iu.edu, lester@ncsu.edu

Abstract

In computer-supported collaborative learning environments, analyzing student dialogue is essential for understanding collaborative problem-solving behaviors and supporting effective learning. Prior work often treats all dialogue interactions uniformly, failing to capture how specific dialogue interaction differentially impact learning experiences and outcomes. To address this limitation, we introduce a dialogue-based learning analytics framework that integrates weighted temporal clustering of dialogue with large language model-based interpretation. Our framework identifies student interaction patterns most predictive of group learning gains and uses these insights to enable early prediction of learning outcomes and generate pedagogically meaningful interpretations. We evaluate our framework on collaborative dialogue from middle school students engaged in a collaborative game-based learning environment. Our results show that our framework achieves 83.1% accuracy in learning outcome prediction. In addition, expert evaluations and case studies demonstrate that the identified weighted dialogue patterns reflect key collaborative problem-solving behaviors recognized as important in collaborative learning. By surfacing high-impact interaction patterns and enabling prioritized interpretation generation, our framework provides a promising approach for accurately analyzing students' collaborative dialogue.

Introduction

In computer-supported collaborative learning (CSCL) environments, teachers play a pivotal role in guiding students toward successful learning. Learning analytics tools, such as teacher dashboards, can enhance teachers' situational awareness and support informed decision-making to better assist their students (Rodríguez-Triana et al. 2018; Van Leeuwen et al. 2014). In particular, collaborative problem solving in CSCL involves complex, ill-structured, computer-mediated activities (Jeong and Hmelo-Silver 2016; Saleh et al. 2019), which benefit from tools helping teachers monitor student engagement and adapt facilitation strategies accordingly. However, simply providing information is unlikely to lead to effective facilitation and may even increase teachers' cognitive load during instructional decisions. Prior research has shown that both educators and learners often struggle to meaningfully interpret analytics feedback or translate it into pedagogical action (Prieto et al. 2017; Sharples 2013). As Tsai (2022) suggests, feedback literacy, i.e., the ability to understand and act on analytics outputs, is essential for realizing the benefits of learning analytics. Similarly, a large-scale study with over 1,600 participants in Germany revealed significant gaps between expectations and actual understanding of learning analytics among students and teachers, underscoring the need for more transparent and actionable feedback mechanisms (Fritz et al. 2024). Hence, there is a need to communicate the results of collaborative problem-solving analysis in ways that are transparent, actionable, and aligned with pedagogical goals.

A promising direction in collaborative problem-solving analytics is to develop automated methods that detect meaningful collaborative problem-solving dialogue and translate its analysis into human-interpretable insights that can inform timely pedagogical action. Recent advances in large language models (LLMs) have made it feasible to automatically interpret student dialogue (Carpenter et al. 2020; Gupta et al. 2023; Kim et al. 2024). For example, previous LLM-enhanced dialogue interpretation frameworks have integrated temporal clustering with LLM-generated explanations to identify interpretable patterns of collaboration and assess their quality (Kim et al. 2025). However, these approaches typically assign equal weight to all detected dialogue patterns when assessing group performance or generating feedback. This uniform treatment risks overlooking the nuanced impact that different types of dialogue interaction can have on learning outcomes.

To address this limitation, we propose a Dialogue Interpretation-based Learning Analytics (DILA) framework that integrates data-driven weighting of temporally clustered dialogue patterns with large language model-based interpretation. By estimating the statistical association between dialogue cluster with learning gains, DILA (1) improves early prediction of group learning outcomes and (2) prioritizes high-impact collaborative problem-solving behaviors in its interpretations. Our contributions are threefold:

- **Weighted dialogue pattern analysis**: A statistical weighting approach that emphasizes patterns most strongly linked to learning gains.

- **Enhanced predictive capability**: Integration of weighting into LLM-based interpretations for more accurate early prediction of group learning outcomes.
- **Actionable interpretation generation**: Weighted interpretations that highlight pedagogically meaningful collaborative problem-solving behaviors, enabling targeted teacher support.

In evaluations using middle school students' collaborative dialogue collected from interactions with a game-based collaborative learning environment, DILA achieved 83.1% accuracy in learning outcomes prediction. By highlighting high-impact dialogue patterns and producing prioritized interpretations, DILA bridges the gap between automated dialogue analysis and actionable classroom interventions.

Related Work

Collaborative problem solving is a process where students integrate their knowledge, skills, and endeavors in collectively solving shared problems (Graesser et al. 2018; OECD 2017). As an essential skill of the 21st century, this competency has been the focus of several theoretical frameworks, particularly within science domains in CSCL environments. Specifically, Liu et al. (2016) introduced a discursive framework for collaboration, in which students' practices are observed through face-to-face conversations and text-mediated communication. The framework involves four categories: (a) sharing ideas, (b) negotiating ideas, (c) regulating problem-solving, and (d) maintaining communication. This framework has been used to map students' collaborative discourse practices and to analyze and understand patterns of their engagement in collaborative problem-solving tasks (von Davier et al. 2017).

Despite advances in the analysis of collaborative problem-solving, it remains critical to present results in a meaningful way that enables teachers to make informed decisions and effectively support students in CSCL settings. The use of learning analytics dashboards as technological support does not necessarily guarantee enhanced situational awareness or informed decision-making. Specifically, Prieto et al. (2017) and Sharples (2013) reported that merely mirroring information about how students engage in collaborative processes, without referencing a desired model or standard for interpreting learning situations, can increase teachers' workload and cognitive burden, potentially limiting their ability to support students. Given the inherent complexity of collaborative problem solving, teachers benefit most from information that clarifies what to attend to and prioritize (Fritz et al. 2024). Grounding this information in a reference model for collaborative practices—rather than simply presenting analytics outcomes—can better support more effective decision making and facilitation (Tsai 2022).

Advances in natural language processing, culminating in large language models (LLMs), have expanded CSCL dialogue analysis from outcome prediction (Gupta et al. 2023) and behavior detection (Carpenter et al. 2020) to richer interpretive tasks such as dialogue act recognition (Kim et al. 2024; Pande et al. 2023) and relational analysis (Chan et al. 2024). In parallel, temporal modeling techniques have been applied to better capture the evolving nature of collaborative interactions. These approaches range from multimodal frameworks that break down complex interactions into structured stages (Ouyang, Zhang, and Graesser 2022), to statistical models such as vector autoregressive methods for identifying time-dependent relationships (Chen, Fiorella, and Nye 2022). Other techniques include sequence and time-series clustering to reveal recurring behavioral patterns (Yang et al. 2023), as well as dialogue transition analysis for mapping shifts between different conversational states (D'Mello, Person, and Lehman 2010).

Recent studies have explored combining LLM-generated interpretations with data-driven refinement. For example, ClickSight uses LLMs to interpret student clickstream behaviors, refining outputs through expert rubrics and self-improvement prompts (Radmehr et al. 2024). Similarly, LLM-based summaries from clustered multimodal reading data (Davalos et al. 2025) and from temporally clustered dialogue patterns (Kim et al. 2025) have been validated and improved based on educator feedback. These approaches demonstrate the value of integrating LLMs with outcome-aware adjustments. Building on this, our work statistically weights LLM-interpreted dialogue clusters by their predictive association with learning outcomes, thereby enhancing interpretability and pedagogical relevance.

Methods

As a baseline, dialogue utterances were embedded and clustered using Dynamic Time Warping (DTW) to capture temporal dynamics. Each cluster was interpreted by an LLM to generate natural language summaries of collaborative problem-solving behaviors. Cluster quality was labeled using (1) LLM-based interpretation of cluster summaries and (2) statistical comparison of cluster occurrences between higher- and lower-performing groups, categorized by normalized group learning gains relative to the overall average. Learning outcomes were then predicted based on the average cluster quality of dialogue segments (Kim et al. 2025).

Dialogue Interpretation-based Learning Analytics

Extending from this baseline, we introduce the Dialogue Interpretation-based Learning Analytics (DILA) framework, which incorporates a data-driven weighting mechanism that assigns importance scores to each dialogue cluster based on its statistical association with learning outcomes, enabling more effective prioritization of high-impact collaborative behaviors in interpretation and teacher feedback generation.

Cluster Weights. To identify and interpret behaviors associated with learning outcomes, we first generate k dialogue clusters through temporal clustering of dialogue sequences using DTW. Next, to reflect its statistical relevance to group learning outcome, we compute a normalized cluster weight (r_i) for each dialogue cluster i. Specifically, we define r_i as the weighted difference of the performance ratio in frequency between higher- and lower-performing groups:

$$r_i = \frac{g|h_i - l_i|}{h_i + l_i} + 1 \tag{1}$$

where h_i and l_i denote the number of higher- and lower-performing groups, respectively, associated with dialogue segments in i-th cluster. The term g is a task-specific weight that amplifies differences when the ratio is small—cases that would otherwise have minimal influence—to enhance discriminative power. The absolute difference is normalized by the total frequency to account for the reliability of the distribution, and we add 1 to avoid zero contribution ($r_i \geq 1$). The larger the value of r_i, the greater the cluster's impact on the learning outcome. Note that r_i is applied only within Equations 2 and 3, not independently of cluster quality scores.

Learning Outcome Prediction. To compute a quality score for a dialogue segment that captures both the assigned cluster labels and their relative influence on learning outcomes, we propose an importance-weighted averaging approach. Each cluster's quality is assessed by an LLM based on its description using scoring criteria defined by educational experts. Each cluster is assigned a cluster score $s \in \{0, 1, 2\}$, representing *emerging*, *developing*, and *exemplary*, respectively.

Dialogue segments are generated using a sliding window (shifted by one utterance) so that each utterance appears in multiple overlapping segments, ensuring a cluster ID can be assigned to every utterance. For a segment of length m, each utterance is assigned a cluster ID using a pre-trained clustering algorithm. These m cluster IDs are then mapped to the corresponding score s and normalized cluster weights, r_i (Equation 1). An utterance-level weighted score is computed as the product of its cluster weight (*r*) and cluster score (*s*), normalized by the total cluster weights. The cumulative weighted score of the dialogue segment, w, is calculated as:

$$w = \frac{\sum_{u=1}^{m} r_u \cdot s_u}{\sum_{u=1}^{m} r_u} \tag{2}$$

where *m* is the number of utterances in that segment.

This weighting ensures that clusters with higher importance have a proportionally larger effect on the final score. As a result, the overall assessment reflects the contribution of the most pedagogically significant discourse patterns rather than treating all clusters equally in the segment. For all dialogue segments in the training data, we calculate w and set the median as the classification threshold; segments with scores above the threshold are labeled high-performing, and those below are labeled as low-performing.

Behavior-to-Learning Gain Mapping. Dialogue clusters derived from bottom-up clustering of dialogue sequences do not correspond directly to expert-informed behavior categories for collaborative problem solving, such as *insufficient participation* or *superficial negotiation* (Table 2). A single cluster can exhibit multiple behaviors; for example, cluster 1 can show insufficient negotiation alongside superficial reasoning. Since cluster weights alone provide limited guidance for educators, it is important to compute weights at the level of these expert-defined behavior categories.

To identify behaviors most strongly associated with learning outcomes, we use an LLM to map cluster descriptions to the behavior categories. For each behavior category, the LLM computes the frequency of observed behaviors across the top $p\%$ of clusters with the highest normalized cluster weight, r_i (Equation 1), where higher frequencies indicate stronger association. In addition, the LLM extracts expressions within each cluster description that serve as behavior indicators. For example, for *insufficient participation*, it identified indicators such as 'contribute less', 'passive', 'disengaged', 'silent', and 'uneven participation' (Table 2). This process produces behavior weights that reflect the relative importance of each behavior for learning and provide meaningful guidance on which behaviors to prioritize for monitoring or pedagogical feedback.

Dialogue Segment Interpretation. For each cluster i in a segment, its importance is computed as:

$$\text{ClusterImportance}(i) = f_i \times r_i \tag{3}$$

where f_i is the cluster's frequency within the segment and r_i is its normalized weight. For example, given a dialogue segment consisting of 30 utterances, if cluster 1 appears 6 times in a 30-utterance segment ($f_1 = 6/30 = 0.2$) and has a normalized weight of 1.5, its importance 0.3.

Clusters are ranked by importance, and the top $p\%$ ones are selected to represent the dialogue segment. Based on empirical observations, we set $p = 30\%$. This process highlights the most salient behavioral patterns by considering both their frequency and pedagogical significance. Arranging these clusters in temporal order reveals the primary dialogue flow trajectories, illustrating how collaboration dynamics evolve over time. Finally, the descriptions of these key clusters are concatenated sequentially to generate an interpretable summary of the corresponding dialogue segment.

Evaluation

Our study uses collaborative dialogue data collected from a collaborative game-based learning environment, ECOJOURNEYS, designed to support middle school students in developing life science understanding and collaborative problem-solving skills (Saleh et al. 2019). In this environment, teams of 3 to 4 middle school students investigated the cause of fish illness in a science-based virtual world, completing structured tasks that prompt collaboration through in-person and in-game dialogue.

The dataset comprises 7,371 utterances—both spoken and typed—collected from 67 students working in 17 collaborative groups. These interactions were captured through video recordings and chat logs during the gameplay. On average, each group generated approximately 434 utterances (SD=237.2), with each quest yielding around 142 utterances (SD=44.3). A paired t-test comparing students' pre- and post-assessment scores revealed a statistically significant improvement ($t(66)$=3.83, p=0.0003), indicating that the learning environment effectively facilitated science learning.

We encode each dialogue segment into a vector representation using a sentence-level transformer-based encoder, MiniLM (384 dimensions) (Wang et al. 2020). DTW is applied to obtain 10 representative clusters from sequences of 20 utterances. For interpretation, we use 30 samples per batch, based on optimal hyperparameters (Kim et al. 2025).

Method	Utterance Window	10	20	30	50	100	200	All	Mean	Diff. (%)
Prior work	LSTM	58.4	50.4	**56.2**	58.2	56.6	**59.5**	**54.9**	56.3	0
	LLM-Expert	**64.7**	**64.7**	52.9	**58.8**	**58.8**	58.8	52.9	**58.8**	**2.5**
	Prior-Base	51.7	45.0	45.0	51.7	45.0	45.0	45.0	46.9	-9.4
	Prior-CPS	61.7	*61.7	55.0	55.0	*55.0	55.0	*55.0	56.9	0.6
	Prior-Expert	*71.7	*71.7	*65.0	*71.7	*65.0	***78.3**	***78.3**	*71.7	15.4
	Prior-CR	***78.3**	***78.3**	***78.3**	**73.3**	***73.3**	73.3	73.3	***75.4**	**19.1**
Proposed	Prior-Base-t	58.3	58.3	58.3	58.3	53.3	53.3	53.3	56.2	-0.1
	Prior-CPS-t	*68.3	*68.3	*68.3	61.7	61.7	61.7	61.7	64.5	8.2
	Prior-Expert-t	***78.3**	***78.3**	***78.3**	*71.7	*71.7	***78.3**	***78.3**	*76.4	15.4
	Prior-CR-t	***78.3**	***78.3**	***78.3**	***78.3**	***78.3**	***78.3**	***78.3**	***78.3**	**22.0**
	DILA-Base	58.3	58.3	53.3	53.3	53.3	53.3	53.3	54.8	-1.5
	DILA-CPS	*68.3	*68.3	*68.3	*68.3	*68.3	*68.3	*68.3	*68.3	12.0
	DILA-Expert	***85.0**	***85.0**	**85.0**	***85.0**	***85.0**	***78.3**	***78.3**	***83.1**	**26.8**
	DILA-CR	*78.3	*78.3	*78.3	*78.3	*78.3	***78.3**	***78.3**	*78.3	22.0

Table 1: The average accuracy of learning outcome early prediction by increasing utterances within a collaborative problem-solving task. Bolded scores represent the best performance within each method group; underlined scores highlight the best overall results for each utterance window. * indicates a significant difference from the baseline LSTM based on a Wilcoxon rank sum test ($p < 0.05$).

Compared Methods. We compare our proposed methods against a range of baselines, including both traditional classifiers and prior collaborative problem-solving (CPS) dialogue analysis frameworks. Our first baseline, the Long Short-Term Memory (LSTM) model, serves as a sequential deep learning baseline trained directly on early dialogue segments. In contrast, our second baseline, LLM-Expert, represents a zero-shot approach that uses a large language model to infer learning outcomes without clustering. Our third baseline, the variants from the prior work (Prior-Base, Prior-CPS, Prior-Expert, and Prior-CR) combine temporal clustering with LLM-generated interpretations under varying levels of collaborative problem-solving framework knowledge and expert guidance. Prior-CR (Cluster Ratio), in particular, leverages statistical frequency differences between high- and low-performing groups to assign quality labels to dialogue clusters. While the prior approaches determined score thresholds using training data from the full dialogue for all utterance windows, our approach improves performance for early prediction by training only on data from utterance windows of the same length, as dialogue patterns change over time. This minor variation is denoted as {*Prior variants*}-*t*, where *t* stands for *truncated*.

Then we compare with variants of DILA:

- **DILA-Base**: Adds data-driven cluster weighting to Prior-Base to reflect statistical relevance to learning outcomes.
- **DILA-CPS**: Extends DILA-Base using the CPS framework definition to improve interpretation quality.
- **DILA-Expert**: Integrate expert-informed CPS behavior prompts with DILA-CPS to enhance pedagogical relevance.
- **DILA-CR**: Integrate cluster-ratio labeling with DILA weighting to enable outcome-driven prediction.

The DILA framework relies on truncated data for the early utterance window.

Learning Outcome Prediction

To evaluate the effectiveness of our proposed DILA framework, we conducted early prediction of group learning outcome across varying lengths of early utterance windows (n=10, 20, 30, 50, 100, 200, All). Table 1 presents learning-outcome prediction accuracy across utterance windows ranging from the first 10 utterances to the entire dialogue. The baseline LSTM attained a mean accuracy of 56.3%, and the zero-shot LLM-Expert improved slightly to 58.8%. Within the original Prior approaches, Prior-Expert (71.7%) and Prior-CR (75.4%) outperformed Prior-Base and Prior-CPS, confirming the value of expert-guided interpretation and outcome-driven labeling. Training the prior approaches' variants on truncated dialogue segments (-*t*) yielded additional gains across most utterance windows, indicating that models benefit from being optimized for limited input contexts. This suggests that dialogue evolves differently across the early and later phases of the quests.

Overall, the proposed DILA models consistently outperformed baselines across all window sizes. DILA-Expert achieved the highest mean accuracy (83.1%), reflecting a 26.8% absolute improvement over LSTM, and preserved strong performance even with only the first 10 utterances (85.0%). DILA-CR matched the performance of Prior-CR-t (78.3%), suggesting that incorporating weighted clusters on top of an already data-driven labeling approach (based on Cluster Ratio) did not yield further improvements. This indicates that when learning-relevant patterns are already emphasized through outcome-informed labeling, additional weighting may offer limited benefit. Collectively, these findings demonstrate that DILA delivers robust, early, and context-wide prediction while enabling more focused, learning-relevant feedback.

Expert-Informed Behaviors		Description	Indicators	Freq.
Negative Category	**Insufficient Participation**	Limited utterances that prevent meaningful engagement (Most frequently noted, reflecting unequal engagement)	contribute less, passive, disengaged, silent, uneven participation	**100%**
	Superficial Negotiation	Arguments that lack evidence or reasoning (Frequently hinders effective consensus)	lack of evidence, shallow, unsupported, superficial, repetitive	**80%**
	Wheel-Spinning Conversations	Repetitive or off-topic talk with unclear goals (Moderate presence, often with frustration or confusion)	off-topic, confusion, unclear, disorganized talk	40%
	Expedited Task	Prioritizing speed over quality dialogue completion	focus on task completion (Rarely emphasized directly)	7%
Positive Category	Comprehensive Contribution	Active discussion of diverse task aspects (Active but uneven involvement is often noted)	active discussion, multiple ideas, sharing resources	53%
	Divergent Idea Sharing	Exchange of diverse and relevant ideas (Occasionally present, but rarely emphasized)	different views, perspectives	27%
	Evidence-Based Negotiation	Reasoned dialogue supported by evidence. (Impactful in presence or absence, despite inconsistent use)	reasoning, support, evidence	**67%**
	Socially-Shared Regulation	Group-level planning, strategies, and reflection (Often mentioned as inconsistent or emerging)	goal setting, reflection, regulation, monitoring	**60%**

Table 2: Importance of 8 expert-informed collaborative problem-solving behaviors from the top 30% of weighted clusters.

Analysis of Weighted Behaviors

To identify which collaborative problem-solving behaviors most strongly influence group learning outcomes, we analyzed the top 30% of clusters that exhibited the largest frequency differences between higher- and lower-performing groups. From the 5-fold cross-validation dataset, we obtained a total of 50 clusters (10 from the training data of each fold) and selected the 15 clusters with the highest weights (30%). We then examined the descriptions of these clusters to assess the extent to which behaviors were observed.

Table 2 presents the frequency of eight expert-informed collaborative problem-solving behaviors within these high-weight clusters, along with their corresponding descriptions. Among the eight behaviors, three were most prominently associated with group learning outcomes and are highlighted in bold. *Insufficient participation* was present in all clusters, typically marked by uneven contributions and disengaged members, which likely hindered collaborative reasoning. The second most frequent behavior was *negotiation*, a composite category reflecting both superficial negotiation and the frequent absence of evidence-based negotiation. These features were observed in 80% and 67% of the high-weight clusters, respectively, and represent a key obstacle to critical thinking and consensus-building. Lastly, *socially shared regulation* appeared in 60% of the high-weight clusters, particularly those associated with higher-performing groups. In these cases, students more consistently engaged in collective planning, goal setting, and monitoring of group processes. These findings suggest that the quality of collaborative problem solving is shaped not only by the depth of cognitive engagement, such as effective negotiation, but also by how regulatory efforts are distributed and coordinated within groups.

Analysis of Dialogue Segments

We demonstrate the interpretation process using student dialogues and examine how weighted and unweighted mechanisms shape the resulting insights. By mapping sequences of dialogue utterances to cluster IDs and linking these IDs to pre-generated cluster interpretations, our approach facilitates collaborative problem-solving analysis and dialogue summarization while minimizing reliance on real-time LLM processing. To analyze the temporal evolution of collaborative problem solving, we first converted each student interaction into its corresponding cluster ID, producing a sequential representation of dialogue behaviors. This representation enabled us to reconstruct each group's behavioral trajectory, revealing whether students advanced toward deeper reasoning and consensus or became caught in cycles of confusion. Based on these temporal patterns, we generated natural language summaries tailored to each group, offering constructive insights into their communication, negotiation, and regulation strategies.

The following example is drawn from actual student conversations during a collaborative game activity, shown on the next page. We extracted the first 30 utterances and assigned cluster IDs to each utterance. To provide a concise overview, we summarized the main flow of the dialogue using a simple LLM prompt, reducing the sequence to representative clusters (e.g., [2, 3, 5, 7, 4, 1]). Based on this reduced sequence, we generated a narrative interpretation, presented as the *Unweighted* version below. Next, we applied our weighting mechanism by computing the weighted sum of each cluster's statistical weight and its frequency within the dialogue (Equation 3). This analysis identified the top three most influential clusters (e.g., [2, 5, 7]). Using only these clusters, we produced a more targeted narrative interpretation, pre-

sented as the *Weighted* version, which highlights patterns more strongly associated with students' learning outcomes.

Sample Dialogue (Utterance Window: 30)

Student 1: so i am kinda leading towards thinking its alge.
Student 1: ok. // Student 1: no.
Student 2: Yall we need to work together on this.
Student 1: cilabrants bacteria.
Student 3: its c. // Student 3: no d. // Student 3: d.
Student 1: were like the slowest team.
Student 2: What do we do next.
Student 4: Does anyone agree with my idea that.. ?
Student 4: As I was playing, I learned that..
Student 4: My reason is..
Student 2: What answer are we going with?
Student 1: so what are we thinking.
Student 3: the mor dead organic matter, the more cloudy the water is, the more .
Student 1: so whats the answer.
Student 2: Which one is that. // Student 2: I think B.
Student 1: i do d. // Student 1: i am thinking b.
Student 3: d again. // Student 3: a.
Student 1: there is another one. // Student 1: a.
Student 2: Its a.
Student 1: a. // Student 1: increases.
Student 1: Bacteria is making the wate cloudy.
Student 3: dead organic matter increases causing the water to get mor and more cloudy.

Examples of Dialogue Summaries

(A) *Unweighted*: The group's collaborative behavior fluctuates between moments of engagement and frustration, marked by varied participation and inconsistent reasoning that weaken cohesion. Despite some respectful communication, unclear ideas and passive involvement lead to confusion and fragile agreements, with shallow negotiation recurring throughout. Dominance by certain members occasionally disrupts balanced dialogue, limiting deeper understanding and consensus-building over time.

(B) *Weighted*: Initially, the group exhibits respectful but imbalanced discussions characterized by shallow negotiation and limited critical reasoning. This shifts toward more engaged but uneven participation, where frustration and inconsistent use of reasoning challenge group cohesion. By the end, active discussions around science topics emerge, though varying engagement and misunderstandings cause superficial negotiation and occasional conflict.

(C) *Combined*: Within the primary focus on respectful yet imbalanced discussions, fluctuating engagement, and superficial negotiation, the group struggles to maintain deep reasoning and critical evaluation consistently over time. Beyond this, the broader sequence reveals repeated cycles of frustration and uneven participation alongside moments of unclear ideas and passive communication that weaken negotiation and regulation. Additionally, dominance by some members and shallow dialogue patterns limit effective sharing and the development of stable, positive collaborative communication throughout the process.

Finally, the *Combined* interpretation incorporates both perspectives: it presents the unweighted cluster sequence, while highlighting the top three most influential clusters from the weighting process. The resulting narrative is designed to balance a holistic view of the dialogue with emphasis on its most impactful behaviors.

Human Expert Analysis. Two educational researchers as coders evaluated summaries generated by three different weighting approaches, *Unweighted* (Summary A), *Weighted* (Summary B), and *Combined* (Summary C), to determine which would be most useful for teachers. Each coder received three anonymized discussion segments from different student groups (each with four members). For each segment, the coders were given three summaries, one per weighting approach, with the model identities concealed as Summary A, B, or C, along with three evaluation questions:

- Estimate the group's learning performance (higher- or lower-performing).
- Rate the group's collaborative problem-solving participation as *exemplary*, *developing*, or *emerging*.
- Which description is most focused and useful as meaningful feedback for students and teachers? Why?

The coders then reviewed the original dialogue segments alongside the corresponding summaries using the collaborative problem-solving framework (Liu et al. 2016) to identify which summaries best captured meaningful student engagement and could help teachers make informed facilitation decisions. Finally, the coders' ratings based on the summaries were compared against actual learning performance and engagement evident in the original dialogue transcripts to assess which summaries most accurately represented students' learning. The responses of the two coders were compared to assess the consistency of their evaluations.

Findings. Using the generated summaries, the two coders assigned levels of learning performance and collaborative problem-solving engagement to evaluate whether the summaries enabled accurate assessment. The two coders showed perfect agreement, consistently rating all three groups as lower-performing across all nine summaries. This suggests that the summaries alone failed to capture meaningful differences in students' learning performance, despite two groups actually demonstrating higher learning outcomes. Although estimating group learning performance from a single 30-utterance segment is challenging, these results indicate that the LLM-generated interpretations alone may not provide sufficient information for accurate assessment.

In terms of collaborative problem-solving engagement, the coders reached about 89% agreement based on the summaries. The sole disagreement occurred for Summary B for Group 1, where one of the coders rated the engagement as *emerging*, and the other as *developing*. For Summaries A and C for Group 1, both rated the engagement as *developing*. For Groups 2 and 3, the coders assigned *emerging* to Summaries A and C, but *developing* to Summary B. Among the three, Summary B most closely aligns with the actual engagement levels observed in the original discourse, whereas Summaries A and C tended to underestimate engagement.

More importantly, the coders showed 100% agreement in selecting the most helpful summary for teachers. They chose Summary C for Group 1 and Summaries B for Groups 2 and 3. For Group 1, Summary C was considered to provide the most comprehensive information about students' engagement in collaborative problem solving and accurately captured key events, making it the most useful for teachers' facilitation decisions among the options. For Groups 2 and 3, Summary B was preferred as it offered more balanced insights, reflecting both positive and negative aspects of collaborative discourse more faithfully. In contrast, Summaries A and C tended to overemphasize negative aspects, reducing their accuracy and effectiveness in supporting teachers' informed decision-making. For instance, Summaries A and C described Group 2's discussion as *chaotic* and marked by multiple *interruptions*. However, upon reviewing the original discourse, the coders did not perceive the students' engagement as chaotic or disruptive. This pattern reveals potential limitations in Summaries A and C. Summary A, lacking explicit guidance on which aspects of collaborative problem solving are most pedagogically meaningful, may have led the LLM to select arbitrary information within the word limit, leading to negative bias. Summary C combines a full-view approach with partial weighting, which can overemphasize certain behaviors. When these behaviors are absent in a discourse segment, their absence may seem disproportionately important. This overemphasis can mask other positive behaviors that occur less frequently or carry lower weight, reducing the overall balance and informativeness of the summary. Consequently, such biased summaries may misrepresent students' collaborative engagement, hindering teachers' accurate assessment and facilitation decisions.

Discussion

A key strength of the proposed weighted clustering approach lies in its ability to highlight dialogue patterns most predictive of learning outcomes and collaborative problem-solving engagement, as reflected in the expert evaluations. This enables more targeted and timely instructional interventions, especially in dynamic classrooms where early detection of unproductive interactions is essential. Beyond prediction, weighting specific dialogue patterns helps educators identify critical moments for impactful intervention. This prioritization could facilitate more efficient use of instructional time and support adaptive feedback linked to pedagogically meaningful categories, enhancing both teacher confidence and student reflection. Furthermore, this flexibility allows the framework to be adapted to different educational and contexts, broadening its applicability.

However, the weighting approach carries risks. Overemphasizing certain behaviors can lead to disproportionate focus on their absence, potentially undervaluing alternative but effective collaboration styles. Such bias can mislead teachers, hindering accurate understanding and informed facilitation. To mitigate these risks, several strategies are advisable. First, combining weighted and unweighted analyses has shown limited impact, highlighting the need for further exploration to capture less-weighted yet valuable behaviors. Second, involving educators in defining and updating weighting criteria maintains alignment with evolving pedagogical goals. Third, implementing adaptable weighting schemes based on data and user feedback can reduce bias toward specific patterns. Finally, delivering balanced feedback that highlights both strengths and areas for improvement is crucial for supporting a holistic view of collaboration.

The expert evaluation revealed challenges in accurately estimating learning performance from limited dialogue segments, suggesting that LLM-generated summaries may not fully capture all information necessary for assessment. Regarding collaborative problem-solving engagement, weighted summaries generally aligned better with the actual discourse, although perfect agreement remained elusive. This underscores the complexity of capturing dynamic and multifaceted collaborative behaviors through automated summarization. Notably, the coders' preferences varied by group; they favored summaries that balanced positive and negative aspects and provided comprehensive, nuanced interpretations. This highlights the importance of interpretive richness in feedback to support teacher decision making. Future work should explore integrating longer-term dialogue data and multiple indicators to enhance assessment validity.

Overall, these findings highlight the necessity of thoughtfully integrating dialogue-based learning analytics into collaborative learning environments. To realize the potential of learning analytics in supporting collaborative problem solving, it is essential to empower teachers and students to critically engage with AI-generated interpretations and adapt feedback to their specific contexts. Such human-centered collaboration will foster more effective and meaningful learning experiences.

Conclusion

We introduced the Dialogue Interpretation-based Learning Analytics (DILA) framework that integrates weighted temporal clustering with LLM-based interpretation to analyze collaborative problem-solving dialogue in game-based learning environments. By estimating the statistical association between interaction patterns and learning outcomes, DILA achieved high early-stage predictive accuracy and provided interpretable, pedagogically meaningful feedback. The weighted clustering approach improved prediction and highlighted key dialogue features, offering actionable insights for teachers and learners. The findings also highlight challenges, including assessing learning performance from limited dialogue and the risk of overemphasizing certain behaviors. Addressing these requires flexible weighting schemes and educator involvement. Future work will explore alternative weighting criteria, hyperparameter optimization, and broader applications to support adaptive scaffolding and reflective learning.

Acknowledgments

This research was supported by funding from the National Science Foundation (NSF) under Grant DRL-2112635. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

Carpenter, D.; Emerson, A.; Mott, B. W.; Saleh, A.; Glazewski, K. D.; Hmelo-Silver, C. E.; and Lester, J. C. 2020. Detecting Off-Task Behavior from Student Dialogue in Game-Based Collaborative Learning. In *Proceedings of the 21st International Conference on Artificial Intelligence in Education, Part I*, 55–66. Springer-Verlag.

Chan, C.; Jiayang, C.; Wang, W.; Jiang, Y.; Fang, T.; Liu, X.; and Song, Y. 2024. Exploring the Potential of ChatGPT on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations. In Graham, Y.; and Purver, M., eds., *Findings of the Association for Computational Linguistics: EACL 2024*, 684–721. Association for Computational Linguistics.

Chen, X.; Fiorella, L.; and Nye, B. D. 2022. Exploring Time-Dependent Interactions in Collaborative Problem Solving Using Vector Autoregressive Modeling. In *Proceedings of the 15th International Conference on Educational Data Mining (EDM)*, 348–359.

Davalos, E.; Zhang, Y.; Srivastava, N.; Salas, J. A.; McFadden, S.; Cho, S.; Biswas, G.; and Goodwin, A. 2025. LLMs as Educational Analysts: Transforming Multimodal Data Traces into Actionable Reading Assessment Reports. *arXiv preprint arXiv:2503.02099*.

D'Mello, S. K.; Person, N.; and Lehman, B. 2010. Mining Collaborative Patterns in Tutorial Dialogues. *Journal of Educational Data Mining*, 2(1): 1–37.

Fritz, B.; Kube, D.; Scherer, S.; and Drachsler, H. 2024. Learning Analytics in Higher Education: Exploring Students and Teachers Expectations in Germany. *arXiv preprint arXiv:2401.11981*.

Graesser, A. C.; Fiore, S. M.; Greiff, S.; Andrews-Todd, J.; Foltz, P. W.; and Hesse, F. W. 2018. Advancing the Science of Collaborative Problem Solving. *Psychological Science in the Public Interest*, 19(2): 59–92.

Gupta, A.; Carpenter, D.; Min, W.; Mott, B.; Glazewski, K.; Hmelo-Silver, C. E.; and Lester, J. 2023. Enhancing Stealth Assessment in Collaborative Game-Based Learning with Multi-task Learning. In Wang, N.; Rebolledo-Mendez, G.; Matsuda, N.; Santos, O. C.; and Dimitrova, V., eds., *Artificial Intelligence in Education*, 304–315. Springer Nature Switzerland.

Jeong, H.; and Hmelo-Silver, C. E. 2016. Seven Affordances of Computer-Supported Collaborative Learning: How to Support Collaborative Learning? How Can Technologies Help? *Educational Psychologist*, 51(2): 247–265.

Kim, Y. J.; Acosta, H.; Min, W.; Rowe, J.; Mott, B.; Chaturvedi, S.; and Lester, J. 2024. Dual Process Masking for Dialogue Act Recognition. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 15270–15283. Association for Computational Linguistics.

Kim, Y. J.; Hong, D.; Min, W.; Chaturvedi, S.; Hmelo-Silver, C. E.; and Lester, J. 2025. Collaborative Problem-Solving Dialogue Analysis with Interpretable Temporal Clustering. In Cristea, A. I.; Walker, E.; Lu, Y.; Santos, O. C.; and Isotani, S., eds., *Artificial Intelligence in Education, Part III*, 30–44. Springer Nature Switzerland.

Liu, L.; Hao, J.; von Davier, A. A.; Kyllonen, P.; and Zapata-Rivera, J.-D. 2016. A tough nut to crack: Measuring collaborative problem solving. In *Handbook of Research on Technology Tools for Real-World Skill Development*, 344–359.

OECD. 2017. PISA 2015 collaborative problem solving framework.

Ouyang, F.; Zhang, Q.; and Graesser, A. 2022. Modeling Multimodal Collaborative Problem Solving Processes with a Three-Level Analysis Framework. *arXiv preprint arXiv:2210.16059*.

Pande, J.; Min, W.; Spain, R. D.; Saville, J. D.; and Lester, J. 2023. Robust Team Communication Analytics with Transformer-Based Dialogue Modeling. In Wang, N.; Rebolledo-Mendez, G.; Matsuda, N.; Santos, O. C.; and Dimitrova, V., eds., *Artificial Intelligence in Education*, 639–650. Springer Nature Switzerland.

Prieto, L. P.; Sharma, K.; Kidzinski, Ł.; and Dillenbourg, P. 2017. Orchestration Load Indicators and Patterns: In-the-wild Studies Using Mobile Eye-racking. *IEEE Transactions on Learning Technologies*, 11(2): 216–229.

Radmehr, M.; Shved, D.; Güreş, B.; Singla, A.; and Käser, T. 2024. ClickSight: Using Large Language Models to Interpret Student Clickstreams in Online Courses. *arXiv preprint arXiv:2505.15410*.

Rodríguez-Triana, M. J.; Prieto, L. P.; Martínez-Monés, A.; Asensio-Pérez, J. I.; and Dimitriadis, Y. 2018. Monitoring Collaborative Learning Activities: Exploring the Differential Value of Collaborative Flow Patterns for Learning Analytics. In *Proceedings of the 18th International Conference on Advanced Learning Technologies (ICALT)*, 155–159.

Saleh, A.; Hmelo-Silver, C. E.; Glazewski, K. D.; Mott, B.; Chen, Y.; Rowe, J. P.; and Lester, J. C. 2019. Collaborative inquiry play: A design case to frame integration of collaborative problem solving with story-centric games. *Information and Learning Sciences*, 120(9/10): 547–566.

Sharples, M. 2013. Shared Orchestration Within and Beyond the Classroom. *Computers & Education*, 69: 504–506.

Tsai, Y. 2022. Why Feedback Literacy Matters for Learning Analytics. In *Proceedings of the International Conference of the Learning Sciences*, 27–34.

Van Leeuwen, A.; Janssen, J.; Erkens, G.; and Brekelmans, M. 2014. Supporting Teachers in Guiding Collaborating Students: Effects of Learning Analytics in CSCL. *Computers & Education*, 79: 28–39.

von Davier, A. A.; Hao, J.; Liu, L.; and Kyllonen, P. 2017. Interdisciplinary Research Agenda in Support of Assessment of Collaborative Problem Solving: Lessons Learned from Developing a Collaborative Science Assessment Prototype. *Computers in Human Behavior*, 76: 631–640.

Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; and Zhou, M. 2020. MINILM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-trained Transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Curran Associates Inc.

Yang, C.-Y.; Munshi, A.; Saab, N.; et al. 2023. Uncovering Collaborative Behavioral Sequences in Game-Based Learning Using Constraint-Based Pattern Mining. In *Proceedings of the 13th International Learning Analytics & Knowledge Conference (LAK)*, 450–461.